**PARTE V: ANALISIS DE INFORMACION**

*PARTE V*

# ANÁLISIS DE LA INFORMACIÓN

# Capítulo 11. INTRODUCCIÓN AI ANÁLISIS DE INFORMACIÓN

En la última etapa, el proceso de investigación econométrica realiza el análisis de la información provista por la tabla de datos; la que fue planteada en el segundo paso y completada de acuerdo a lo establecido en tercera y cuarta parte. Uno de los objetivos en el planteo de la tabla de datos es que, en este quinto paso, se realice el estudio de la semejanza entre las unidades de observación y la asociación entre las variables. En este capítulo se ven las técnicas apropiadas al análisis de la asociación entre las variables y en el próximo se estudian los métodos para determinar la semejanza entre las unidades de observación. En ambos casos, los métodos para realizar estos estudios difieren de acuerdo a que si sobre las unidades de observación se han realizado mediciones con variables cualitativas o con variables cuantitativas o con ambas. El grado de asociación entre dos variables cuantitativas se establece mediante el coeficiente de correlación. Cuando ambas variables son cualitativas se aplica el estudio de la asociación mediante el análisis de tablas de contingencia. Por último, cuando sobre una variable cuantitativa se desea estudiar la dependencia, o efectos, que sobre su comportamiento tienen modalidades de una variable cualitativa, el método a emplear es el análisis de la varianza.

## 11.1. Correlación de muestras

La correlación mide el grado de asociación entre variables cuantitativas. La correlación muestral refleja la tendencia para que los puntos se agrupen sistemáticamente alrededor de una línea recta que crece o decrece de izquierda a derecha; se observa gráficamente al representar dos variables sobre una gráfica bidimensional denominado Diagrama de Dispersión.

La Figura 11.1 muestra la dispersión de las variables PBI a nivel nacional e Indice de Evolución Económica de Río Cuarto (INEVE), las mismas presentan una asociación estadística positiva. Una correlación positiva, reflejará una tendencia de alto valor para una primera variable que está asociada con un alto valor de una segunda. Una correlación negativa, refleja una asociación entre un valor alto de una primera variable y un valor bajo de una segunda variable.

El coeficiente de correlación muestral se define como,

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1)\, s_x\, s_y}; \quad -1 \le r \le 1$$

El valor del coeficiente de correlación cercano a +1, indica una **asociación positiva perfecta** ente las dos variables; mientras que, si el coeficiente de correlación es cercano a -1, existe una **asociación negativa perfecta**. Un coeficiente de correlación cercano a 0, refleja la ausencia de asociación lineal.

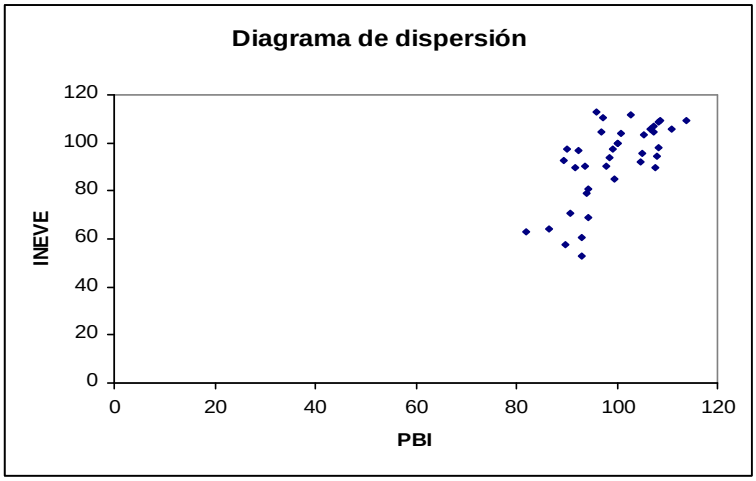


**Figura 11.1 Diagrama de dispersión**

**Ejemplo.** La varianza observada en el indicador local de actividad económica (INEVE) es de 274.55 y la varianza en el indice de PBI 56.01. La covarianza entre ambas variables es de 71.56. ¿Cuál es la correlación entre ambas variables?

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1)\, s_x\, s_y} = \frac{Cov(X,Y)}{\sqrt{S_X^2}\sqrt{S_Y^2}} = \frac{71.56}{\sqrt{274.55}\sqrt{56.01}} = 0.5927$$

## 11.2 Análisis de Tablas de Contingencia

Para estudiar la asociación de variables cuantitativas se utiliza el coeficiente de correlación. Cuando se trabaja con variables cualitativas, en lugar de la correlación, se analiza la independencia de las variables. Dos variables son estadísticamente independientes, si el conocimiento de una no ofrece información sobre la identidad de la otra.

Para llevar un análisis de esta naturaleza se necesita dos variables cualitativas con modalidades mutuamente excluyentes entre sí en cada variable. La cantidad de observaciones que presentan modalidades de las dos variables, en el mismo momento, se disponen en una tabla que se denomina de Contingencia. Esta Tabla de Contingencia reúne la totalidad de coocurrencias que presenta la población para dos variables consideradas.

La Tabla de Contingencia que surge de la observación, censal o muestral, se la denomina Tabla de frecuencias observadas. Tanto las filas como las columnas, brindan información sobre las modalidades de las variables. Por esto es importante conocer el total de observaciones en cada modalidad y el peso que éste total modalidad tiene en el total de observaciones realizadas en la población. A éste último indicador se lo denomina perfil; así se tiene el perfil columna (peso del total modalidad de la columna en el total de observaciones) y el perfil fila (peso del total modalidad de la fila en el total de observaciones).

Los perfiles (fila y columna) surgidos de una tabla de frecuencias observadas permiten construir la tabla de frecuencias esperada bajo situación de independencia. Esta tabla es teórica y se construye con el producto de los perfiles fila y columna y el total de observaciones. Formalmente

$$p_f \, x \, p_c \, x \, n$$

Donde $p_f$ y $p_c$ son los perfiles fila y columna medidos en número de observaciones y $n$ es la cantidad de observaciones.

## El estadístico chi-cuadrado

El estadístico chi-cuadrado, es una medida de la diferencia entre las frecuencias observadas con la frecuencia esperada bajo el supuesto de independencia estadística. El estadístico de chi-cuadrado, se define como:

$$\lambda^2_{(f-1)(c-1)} = \sum \frac{(f_a - f_a^e)^2}{f_a^e}$$

donde, $f_a$ es la frecuencia absoluta observada, $f_a^e$ es la frecuencia absoluta esperada bajo el supuesto de independencia, $f$ es el número de filas, $c$ es el número de columnas y $(f-1)(c-1)$ son los grados de libertad

Si las variables son estadísticamente independientes, el valor ¡$^2$ debería ser relativamente pequeño. Sin embargo, si las variables no son independientes -si están asociadas o relacionadas-, entonces el valor de ¡$^2$ debería ser relativamente grande.

## La prueba de la hipótesis

La **hipótesis nula** asociada con el estadístico muestral chi-cuadrado, es que las dos variables cualitativas son estadísticamente independientes.

La **hipótesis alternativa**, es que las dos variables no son independientes.

La prueba de la hipótesis se basa en el hecho de que el estadístico chi-cuadrado está distribuido como la distribución chi-cuadrado con $(f-1)(c-1)$ grados de libertad, suponiendo que la hipótesis nula es verdadera. O, en otras palabras, dado un nivel de significación $\alpha$ se puede determinar, de acuerdo a los grados de libertad, el valor teórico de Chi- Cuadrado, se lo compara con el valor empírico y si el segundo es superior, se rechaza la hipótesis de independencia.

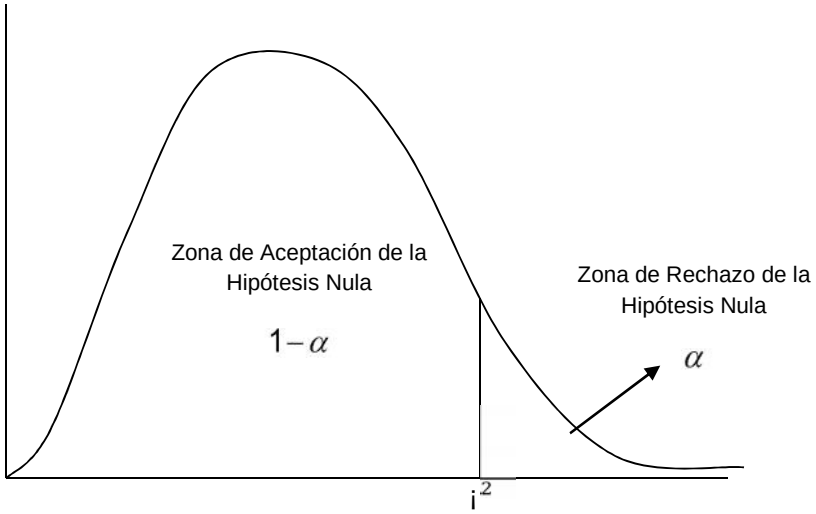


**Figura 11.2 Prueba de la hipótesis de independencia**

Ejemplo. La Figura 11.3 muestra los resultados de una encuesta de 200 asistentes a las funciones de ópera que fueron interrogados acerca de la frecuencia con la que asistían a los conciertos de la orquesta sinfónica en una ciudad vecina. La frecuencia de asistencia, fue dividida en las categorías de frecuente, ocasional y nunca; y se les preguntó si consideraban la ubicación de la sinfónica como conveniente o inconveniente.

| Asistencia a los conciertos de la sinfónica (A) | Ubicación (U) | | Total | Perfil fila $p_f$ |
|---|---|---|---|---|
| | Conveniente | No conveniente | | |
| Frecuente | 22 | 18 | 40 | 0.20 |
| Ocasional | 48 | 52 | 100 | 0.50 |
| Nunca | 10 | 50 | 60 | 0.30 |
| Total | 80 | 120 | 200 | 1.00 |
| Perfil columna ¡2 $p_c$ | 0.40 | 0.60 | 1.00 | |

**Figura 11.3 Asistentes a la Opera – Frecuencia Observada**

La tabulación cruzada resultante, muestra la clasificación porcentual de asistencia en cada categoría de localización. Los totales de las filas, los totales de las columnas y las proporciones ($p_f$ y $p_c$), son tabulados al margen. $p_f$ y $p_c$ son las distribuciones de frecuencias marginales para

las variables respectivas. Por ejemplo, la fila total indica que 80 entrevistados (0.40 de todos los entrevistados), consideraban que la localización era conveniente y 120 (0.60 de todos los entrevistados) que era inconveniente.

Si el experimento se repitiera, un 20% de los resultados mostrarían que la Asistencia a los conciertos de la sinfónica pertenece a la categoría de "frecuentes". La frecuencia absoluta esperada, bajo la condición de independencia estadística, sería de 0.20n, donde n es el número de nuevos experimentos realizados. También se espera, bajo las mismas condiciones que el 40% de los casos tengan una ubicación "conveniente" y el 60% una ubicación "no conveniente".

De igual manera, el número de experimentos que darían como resultado una asistencia frecuente en una ubicación conveniente "se esperaría" que fueran $0{,}4x0{,}2xn$

| Asistencia a los conciertos de la sinfónica (A) | Ubicación (U) | | Total | Perfil fila $p_f$ |
|---|---|---|---|---|
| | Conveniente | No conveniente | | |
| Frecuente | 16 | 24 | 40 | 0.20 |
| Ocasional | 40 | 60 | 100 | 0.50 |
| Nunca | 24 | 36 | 60 | 0.30 |
| Total | 80 | 120 | 200 | 1.00 |
| Perfil columna $p_c$ | 0.40 | 0.60 | 1.00 | |

**Figura 11.4 Asistentes a la Opera – Frecuencia esperada**

El valor empírico para el estadístico chi cuadrado es:

$$i^2_{(f-1)(c-1)} = \sum \frac{(f_a - f_a^e)^2}{f_a^e} =$$

$$= \frac{(22-16)^2}{16} + \frac{(18-24)^2}{24} + \frac{(48-40)^2}{40} + \frac{(52-60)^2}{60} + \frac{(10-24)^2}{24} +$$

$$+ \frac{(50-36)^2}{36} = 20.02$$

El estadístico de chi-cuadrado teórico, para 2 grados de libertad y nivel de confianza de 0.95, es de 5.991. Este valor crítico divide la zona de aceptación de la hipótesis nula de la zona de rechazo.

El valor empírico de 20.02 es mayor al teórico, esto indica que las dos variables pueden no ser estadísticamente independientes; al menos en este caso particular, se encontró una **asociación muestral** entre las dos variables.

## 11.3 Análisis de varianza

Este método se utiliza para comparar las diferencias entre las medias de diferentes grupos, al combinar, en el mismo análisis, variables cuantitativas y variables cualitativas.

El objetivo es conocer si existen diferencias en los valores medios de la variable cuantitativa en cada modalidad de la variable cualitativa, bajo la hipótesis nula de igualdad entre las medias de los distintos grupos. Esta hipótesis se formula:

$$H_0 : «_1 = «_2 = \cdots = «_k$$

La hipótesis alternativa es que las medias no son todas iguales

$$H_1 : \text{no todas las } «_j \text{ son iguales } j = 1, 2, \cdots, k$$

El estadístico de contraste tiene en cuenta las variaciones cuantitativas dentro de cada modalidad y las variaciones cuantitativas entre las modalidades. Formalmente es

$$F = \frac{MSA}{MSW} \sim F_{k-1;n-k}$$

Donde: $MSA$ es la varianza entre grupos; $MS$ es la varianza dentro de los grupos; k es el número de modalidades o grupos en los que particiona la variable cualitativa y n el número total de observaciones.

---

**Ejemplo.** Se cuenta con la calificación a 15 integrantes de un programa técnico. Este programa consta de tres métodos de enseñanza -A, B y C- que desarrollan un nivel determinado de habilidad en diseño auxiliado por computadora; estos métodos fueron asignados en forma aleatoria entre los participantes.

La variable cuantitativa presente en el análisis es la calificación obtenida y la variable cualitativa el método de instrucción que consta de 3 modalidades.

La Figura 11.5 presenta las calificaciones alcanzadas, al término de la unidad de instrucción en cada método, y las calificaciones promedio correspondientes.

En primer lugar se calcula:

$$\boldsymbol{X_j = \sum_{i=1}^{n_k} X_{ij} \quad \forall\, j = A, B, C}$$

$$\boldsymbol{\bar{X}_j = \sum_{i=1}^{n_k} X_{ij} \Big/ n_k \quad \forall\, j = A, B, C}$$

Donde $X_j$ es la suma de todas las calificaciones (en cada grupo); $X_{ij}$ es la calificación promedio obtenida por cada participante; $\bar{X}_j$ es la media de calificaciones en cada modalidad y $n_k$ es la cantidad de observaciones en cada grupo.

| Instrucción | Calificaciones de la prueba por participante | | | | | Calificaciones Totales $X_j$ | Calificaciones Promedio $\bar{X}_j$ |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| A | 86 | 79 | 81 | 70 | 84 | 400 | 80 |
| B | 90 | 76 | 88 | 82 | 89 | 425 | 85 |
| C | 82 | 68 | 73 | 71 | 81 | 375 | 75 |

**Figura 11.5 Calificaciones alcanzadas por los participantes en cada grupo**

El cálculo de la media de todos los elementos de la muestra ($\bar{X}_{..}$) se calcula a partir del doble sumatorio, uno para sumar los elementos dentro de la fila y el otro para sumar los totales entre los grupos.

$$\bar{X}_{..} = \sum_{j=1}^{k}\sum_{i=1}^{n_k} X_{ij} \Big/ k * n_k = \sum_{j=1}^{3}\sum_{i=1}^{5} X_{ij} \Big/ 3 * 5 = \frac{400 + 425 + 375}{15} = 80$$

Para llevar a cabo una prueba de análisis de varianza (ANOVA), se subdivide la variación total ($SST$) en aquellas que pueden atribuirse a la variación entre grupos ($SSA$) y la que se debe a variaciones dentro de los grupos ($SSW$); de modo que

$$SST = SSA + SSW$$

Ahora se calculan estas variaciones:

- variación total ($SST$)

$$\begin{aligned} SST &= \sum_{j=1}^{k}\sum_{i=1}^{n_k}(X_{ij} - \bar{X})^2 = \sum_{j=1}^{3}\sum_{i=1}^{5}(X_{ij} - 80)^2 \\ &= (86-80)^2 + (79-80)^2 + (81-80)^2 + (70-80)^2 + (84-80)^2 \\ &+ (90-80)^2 + (76-80)^2 + (88-80)^2 + (82-80)^2 + (89-80)^2 \\ &+ (82-80)^2 + (68-80)^2 + (73-80)^2 + (71-80)^2 + (81-80)^2 \\ &= 698 \end{aligned}$$

- variación entre grupos ($SSA$)

$$\begin{aligned} SSA &= \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2 = \sum_{j=1}^{3} 5\left((80-80)^2 + (85-80)^2 + (75-80)^2\right) \\ &= 5(0 + 25 + 25) = 50 * 5 = 250 \end{aligned}$$

- variación dentro del grupo ($SSW$)

$$\begin{aligned} SSW &= \sum_{j=1}^{k}\sum_{i=1}^{n_k}(X_{ij} - \bar{X}_j)^2 = \sum_{i=1}^{5}(X_{iA} - 80)^2 + \sum_{i=1}^{5}(X_{iB} - 75)^2 + \sum_{i=1}^{5}(X_{iC} - 85)^2 \\ &= (86-80)^2 + (79-80)^2 + (81-80)^2 + (70-80)^2 + (84-80)^2 \\ &+ (90-85)^2 + (76-85)^2 + (88-85)^2 + (82-85)^2 + (89-85)^2 \\ &+ (82-75)^2 + (68-75)^2 + (73-75)^2 + (71-75)^2 + (81-75)^2 \\ &= 448 \end{aligned}$$

El resultado de los cálculos anteriores se dispone en la Figura 11.6. La hipótesis nula a probar es

$$H_0: \mu_A = \mu_B = \mu_C$$

Al nivel de confianza de 0.95, el estadístico crítico es

$$F_{k-1;n-k} = F_{3-1;15-3} = F_{2;12} = 3{,}89$$

| Fuente de Variación | | Grados de libertad | Varianzas |
|---|---|---|---|
| Entre fila | $SSA = 250$ | $k - 1 = 3 - 1 = 2$ | $MSA = \frac{SSA}{k-1} = \frac{250}{2}$ |
| Dentro de la fila | $SSW = 448$ | $n - k = 15 - 3 = 12$ | $MSW = \frac{SSW}{n-k} = \frac{448}{12}$ |
| Total | $SST = 698$ | $n - 1 = 15 - 1 = 14$ | $MST = \frac{SST}{n-1} = \frac{698}{14}$ |

**Figura 11.6**

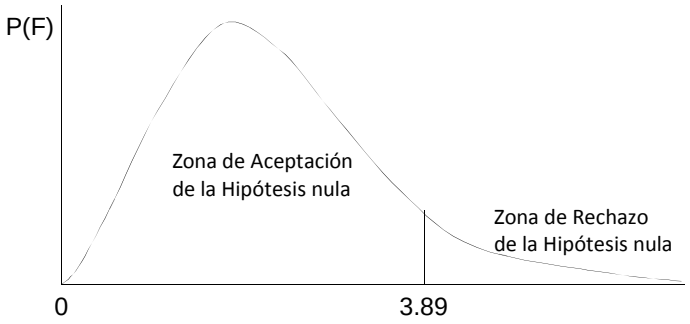


**Figura 11.7 Prueba de la hipótesis de igualdad de medias**

El valor empírico se calcula

$$F^e = \frac{MSA}{MSW} = 3{,}35$$

Este valor es menor que el valor crítico $\left(F^e < F_{2,12}\right)$ por lo que se acepta la hipótesis nula; las calificaciones promedios son iguales entre los diferentes métodos de instrucción.

## CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

### Caso 11.1: Razas de perros

A partir de la tabla de datos generada en el Caso 9.2, seleccione dos variables y pruebe la hipótesis de independencia.

## Problemas

11.1. En el periodo Marzo 1993 y Diciembre 2003, la varianza observada en el consumo de energía eléctrica alcanzó 815151.19 Gwh y la varianza en el PBI 374.07 miles de millones de pesos a valores de 1993. La covarianza entre ambas variables es de 6204.11. ¿Cuál es la correlación entre ambas variables?

11.2. El consumo de gas observado en la Argentina en el periodo Marzo de 1993 y Diciembre de 2003 registró una variación de 510030.89 millones de metros cúbicos y la covarianza con el consumo de energía eléctrica alcanzó, en el mismo periodo 566380.53. Con el dato de varianza en el consumo de energía eléctrica de 815151.19 Gwh, calcule el coeficiente de correlación entre el consumo de Gas y el consumo de energía eléctrica.

11.3. En la tabla de contingencia se muestran las 200 personas que entraron en una tienda de equipos de sonido de acuerdo con el sexo y la edad. Pruebe si las dos variables categóricas son independientes.

| Clientes de la tienda de aparatos de sonido | | | |
|---|---|---|---|
| | Sexo | | |
| Edad | Hombre | Mujer | Total |
| Menor de 30 | 60 | 50 | 110 |
| 30 y más | 80 | 10 | 90 |
| Total | 140 | 60 | 200 |

11.4. En la tabla se muestran las reacciones de los votantes ante un nuevo plan de impuestos sobre bienes raíces, de acuerdo con su afiliación política partidaria. Con estos datos, construya una tabla de frecuencias esperadas suponiendo que no existe relación ante el plan fiscal.

| Reacciones de los votantes ante un nuevo plan de impuestos | | | | |
|---|---|---|---|---|
| | Reacción | | | |
| Afiliación partidista | A favor | Neutral | Se opone | Total |
| Partido A | 120 | 20 | 20 | 160 |
| Partido B | 50 | 30 | 60 | 140 |
| Otro | 50 | 10 | 40 | 100 |
| Total | 220 | 60 | 120 | 400 |

11.5. En la tabla siguiente se presenta la reacción de los estudiantes ante la ampliación de un programa deportivo colegial, de acuerdo con la clase a la que pertenecen. División inferior indica que se trata de un alumno de primer o segundo año y División superior señala que los alumnos se encuentran en el tercer o cuarto año, siendo ésta la división de clase. Pruebe la hipótesis nula de que la posición de clase y la reacción ante el programa deportivo son variables independientes, utilizando el nivel de significación del 5%.

| Reacción de los estudiantes ante el plan deportivo de acuerdo a su generación | | | |
|---|---|---|---|
| | Generación | | |
| Reacción | División inferior | División superior | Total |
| A favor | 20 | 19 | 39 |
| En contra | 10 | 16 | 26 |
| Total | 30 | 35 | 65 |

11.6 Un nutricionista dividió aleatoriamente a 15 ciclistas en tres grupos de 5 cada uno. Los integrantes del primer grupo recibieron suplementos vitamínicos que añadieron a sus corridas durante las tres semanas siguientes. A los del segundo grupo se les indicó que, durante esas tres semanas, tomaran un tipo especial de cereal de grano entero rico en fibra. A los miembros del tercer grupo se les dijo que comieran como hacían normalmente. Transcurrido el periodo indicado, el nutricionista hizo que cada ciclista recorriera una distancia de 6 kilómetros en la que registraron los siguientes tiempos:

| Grupo | Tiempos en el recorrido de 6 kilómetros | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| De las vitaminas | 15,6 | 16,4 | 17,2 | 15,5 | 16,3 |
| Del cereal rico en fibra | 17,1 | 16,3 | 15,8 | 16,4 | 16,0 |
| De control | 15,9 | 17,2 | 16,4 | 15,4 | 16,8 |

Estos datos ¿son consistentes con la hipótesis de que ni las vitaminas ni el cereal rico en fibras afectan a la velocidad de los ciclistas? Utiliza un nivel de significación del 0,05.

## Bibliografía


- **Berenson, Mark y Levine, Daniel.** *Estadística Básica En Administración*. México: Prentice Hall, 1996.
- **____.** *Estadística Para Administración Y Economía. Conceptos Y Aplicaciones.* México: Mc. Graw Hill, 1993.

- **Box, G - Hunter, J - Hunter, W.** *Estadística Para Investigadores. Diseño, Innovación Y Descubrimiento*. Barcelona: Editorial Reverté SA, 2008.
- **Daniel, W.** *Bioestadística, Base Para El Análisis De Las Ciencias De La Salud.* México: Editorial Limussa, 1999.
- **Hildebrand, D. y Lyman Ott R.** *Estadística Aplicada a La Administración Y a La Economía*. Wilmington, USA: Addison Wesley Iberoamericana, 1997.
- **Kazmier, L y Diaz Mata, A.** *Estaística Aplicada a La Administración Y a La Economía*. México: McGraw Hill, 1993.